Project Abstract

The Grammars of Human Behavior
0433136
Yiannis Aloimonos and Ken Nakayama
University of Maryland and Harvard University

Have you ever thought what goes on inside your head when you look at someone performing an action and you understand it? Or how about when I show you a dance and you look at me and you are learning the movements? And what about the case where you think of an action that you did yesterday or that you will do tomorrow? We do these things so effortlessly that we hardly ever think how we accomplish them. These scenarios show that when these behaviors are exhibited, we are relating different action representations to each other. The action representational system is not monolithic, but rather occupies a spectrum of informational structures at hierarchical levels corresponding to different behavior "spaces": (a) the **mechatronic space** used in movement planning and production, i.e when we actually perform an action (b) the **cognitive space**, involving representations for action recognition, analysis, and evaluation; (c) the **visual motion space**, which encodes and organizes visual motion caused by human action; and (d) the **linguistic motion space**, comprised of conceptual/symbolic action encoding.

Our theoretic, computational, and experimental efforts seek to clarify and formally describe both the nature of the representations in these spaces and, crucially, the mapping of representations across spaces. Notably, we explore a candidate action representation, referred to as a **visuomotor** representation, which, in facilitating the understanding of observed actions, may recapitulate and resonate with the actual motor representations used to generate movement. Moreover, we present a promising approach for obtaining this representation from discrete action elements or anchors.

This endeavor spans a number of **research** domains, both **basic** and **applied**, including **Human and Computer Vision** (e.g., automated action recognition in digital video, surveillance, security), **Cognitive and Social Psychology** (e.g., robust social judgments given degraded biological motion), **Kinesiology and Motor Control** (e.g., analysis/modeling/training of movement profiles, as in athletics or pathology/rehabilitation), **Artificial Intelligence and Robotics** (e.g., control of anthropomorphic robots and symbol grounding), and **Computer Science and Animation**.

The Intellectual Merit of our proposed work derives from its principled development and empirical evaluation/refinement of a novel formal theory of the mental representations and processes subserving action understanding and planning; our work provides a compact but powerful and extensible computational approach to the analysis *and* synthesis of complex actions (and action sequences) based on a very small set of atomic postural elements ("keyframes" or "anchors") and the corresponding probabilistic, grammatical rules for their combination. Thus, in a sense, our probabilistic "pose grammar" approach to action representation is similar to state of the art techniques used for speech recognition (e.g., hidden Markov models), but with key postural silhouettes taking the place of phonemes; such augmented transition grammars also

nicely reflect sophisticated new control-theoretic techniques in Robotics for robust anthropomorphic movement.

There are **two promising thrusts** of our work, one **scientific** and the other **technological.** Our studies on human action reveal a structure similar to language, as they both (action and language) have a recognitive and generative aspect (I understand what you say and I can also produce language; I understand with my vision what you do and I can do a similar action). Like in language, human action has nouns, adjectives, verbs, adverbs and prepositions. We are collecting empirical data on thousands of human actions and we are learning a grammar that produces all these actions. We think that this grammar of action will play a fundamental role in understanding the ideas surrounding **Chomsky's Universal Grammar**. The technological thrust comes from software that can automatically recognize human actions in video. Although many researchers work on this problem, our approach is unique in the sense that **we use both visual representations** (that we extract from images) **and motor representations** that we learn from measuring human action.

Project Website

http://www.cs.umd.edu/~karapurk/nsfhsd/index.htm